A Two-Stage Estimation Approach for Bridging the Performance-Complexity Gap in Joint Source Channel Decoding

A Two-Stage Estimation Approach for Bridging the Performance-Complexity Gap in Joint Source-Channel Decoding

Yu-wei Wang†, Elias S.G. Carotti‡, David J. Miller†, Juan Carlos de Martin*

†Dept. of Elec. Eng.
The Pennsylvania State University
Rm. 227-C EEW
University Park, PA 16802
email:millerdj@hyperion.ee.psu.edu

‡Politecnico di Torino
c.so Duca degli Abruzi, 24
10129 Torino, Italy

*IEIIT-CNR
c.so Duca degli Abruzi, 24
10129 Torino, Italy

Abstract — **Joint source-channel decoding techniques capitalize on residual redundancy that typically remains following a source encoding operation. These methods, which include MAP and MMSE-based decoders, estimate the sequence of encoded source symbols based on statistical knowledge of both the channel and the encoded source. Generally, these techniques are based on a Markov model for the quantized source and, thus, on a hidden Markov model for the source-channel tandem. The number of states in the hidden Markov model, and thus the computational and storage complexities, grow *exponentially* with the order (K) of the Markov model, i.e., the complexity order is $O(N^{K+1}T)$, with N the number of source quantization levels and T the length of the data sequence. Thus, to retain implementable complexity, low order models (K = 1,2) are typically used, at the expense of model accuracy. In this work, we propose a method to *bridge* the performance-complexity gap, i.e. to provide solutions that give better performance than a low order decoder while incurring only modest increases in complexity. Our decoding approach consists of two stages: 1) low order JSC decoding, followed by 2) a linear FIR filtering of the JSC decoded signal. The linear filter is chosen to provide an optimal (least squares) estimate of the original source. This approach provides an approximate way to increase the effective order of the decoder, yet while retaining quite manageable complexity. The new approach is demonstrated to significantly improve upon standard MMSE-based JSC decoding performance, both for the case of nonpredictive source coding (e.g. vector quantization) as well as for predictive source coding (DPCM).**

I. Introduction

In [13], for a differential pulse code modulation (DPCM) system, a suboptimal predictor was used, introducing statistical dependencies within the sequence of quantization indices output by the DPCM encoder. The authors proposed to exploit this redundancy at the decoder to improve error resilience in much the same way that controlled redundancy inserted via channel coding is exploited. They developed a joint source-channel (JSC) decoder that finds the most likely sequence of transmitted indices, given a sequence of noisy received indices. Their technique requires that the decoder have access to a simple model for the quantized source as well as a model for the channel. It was demonstrated in [13] that significant error-resilience could be achieved. Following this work, there has been substantial further activity in *decoding based on residual redundancy.*

Both maximum *a posteriori* (MAP) [12] and minimum mean-squared error (MMSE) [8] techniques have been proposed. These JSC decoding techniques have been applied both to non-predictive coding systems (e.g. scalar and vector quantization) [8, 9, 12] as well as to predictive coding systems (e.g. DPCM) [4, 6, 10, 13]. They have also been applied both to time series and to images. One difficulty with these methods, especially when applied to images, is the computational and storage complexity of the decoding.

JSC decoding uses a Markov model for the quantized source. The concatenation of the source and channel is, thus, a *hidden* Markov model, with the quantized source playing the role of the hidden state sequence. The decoder's problem thus boils down to estimating the state sequence, or some function of it (in particular the sequence of quantized values). Dynamic programming/the Viterbi algorithm can be used to identify the most likely state sequence. Alternatively, the Forward/Backward [1] algorithm finds the *a posteriori* state probabilities, which can be used for conditional mean estimation, i.e. for MMSE estimation. Unfortunately, in either case, the decoding complexity increases *exponentially* with the Markov model order K, i.e. the complexity is $O(N^{K+1}T)$, with N the cardinality of the state and T the sequence length. Thus, in the literature, one typically only sees low orders investigated, e.g. K=2 or 3. In the case of an image source, the computational difficulties are only accentuated. In [10] a Markov mesh model was suggested for a DPCM encoded source. Even for the most minimal causal conditioning context, based on one pixel to the left and one pixel above the current pixel of interest, the complexity of exact a posteriori state estimation goes as $O(N^5T)$. In [11], approximate state estimation was performed, with $O(N^3T)$ complexity. Clearly, as the conditioning context increases, exact state estimation becomes practically intractable. Even with high order conditioning contexts, there are ways of keeping the decoding complexity manageable, but at the price of estimation accuracy. One practical solution is to use greedy techniques, e.g. [4], where hard state estimation was performed "one pixel at a step". In this approach, when estimating the state at a current pixel of interest, one uses as conditioning context the decoded values in a causal neighborhood of this pixel. The *storage* complexity of this approach will *still* grow exponentially with conditioning order. However, the computational complexity is now only *linear* in the state cardinality, i.e. $O(NT)$. A related *soft* state estimation technique was proposed in [6]. Here, a posteriori state probabilities are computed "one pixel at a step", using previously computed a posteriori probabilities in the causal neighborhood of the current pixel. However, unlike the "hard" case[4], the decoding complexity for the approach in [6] *still* grows exponentially, in particular $O(N^{K+1}T)$, with K the (causal) conditioning order. There are also some techniques for improving upon hard greedy searches. In [2], an iterative hillclimbing algorithm was proposed with guaranteed improvement over greedy search. In this approach, instead of optimizing "one pixel at a step", one jointly optimizes over an entire row or column at each step, achieved via dynamic programming. Moreover, whereas the greedy search in [4] terminates after one pass over the image, in [2] the solution can be iterated, with multiple row/column sweeps taken over the image, and with the performance guaranteed to improve (in the sense of the defined objective function) with each image pass. The complexity of this method is $O(NTL)$, with L the number of image iterations taken. From the discussion above, we can see that there is a fundamental difference in complexity in the greedy hard and soft estimation cases. Computation grows linearly in N in the hard case and as N^K in the soft case. However, it is also well-known that soft estimation techniques generally outperform their hard counterparts. The reason being that they preserve many hypotheses for the hidden state sequence (and attach associated probabilities), whereas hard state estimation preserves only one such hypotheses. In the present paper, we derive a general decoding technique that aims to capture some of the advantage of higher order soft estimation while still retaining low complexity. We thus suggest an approach which can bridge the performance gap between low order (low complexity) and high order (high complexity) JSC decoding. Our approach is applicable to both non-predictive and predictive coding systems, as will be discussed in the sequel.

II. New JSC Decoding Formulation

Preliminaries

Consider the problem of source coding, with transmission of the encoded information over a noisy channel. The communications system model is shown in Fig. 1. For clarity's sake, we only consider the memoryless channel case here. The source encoder takes $X_t \in \mathcal{R}$ and produces an output index $I_t \in \mathcal{I}$, with $\mathcal{I}$ the quantization index set $\{0, 1, \ldots, N-1\}$. The index is transmitted over a discrete channel, resulting in a possibly corrupted index J_t, from the same set $\mathcal{I}$, according to the (assumed constant) channel transition probabilities $\{P[J_t = j_t | I_t = i_t]\}$. We define the data source, from a discrete-time random process, as the sequence $\underline{X} \equiv (X_0, X_1, \ldots, X_{T-1})$, the encoder's index sequence $\underline{I} \equiv (I_0, I_1, \ldots, I_{T-1})$, and the received sequence, $\underline{J} \equiv (J_0, J_1, \ldots, J_{T-1})$. As in prior work [8, 12, 13], we assume a Markovian model for the sequence of transmitted indices, e.g. in the first order case,

$$\begin{aligned} & P[I_0 = i_0, I_1 = i_1, \ldots, I_t = i_t] \\ &= P[I_0 = i_0] \prod_{l=1}^{t} P[I_l = i_l | I_{l-1} = i_{l-1}], \quad \forall t. \end{aligned} \tag{1}$$

All the probabilities $\{P[J_t = j_t | I_t = i_t]\}$, $\{P[I_0 = i_0]\}$, $\{P[I_t = i_t | I_{t-1} = i_{t-1}]\}$ are assumed known at the decoder. In practice, the source probabilities are estimated based on an encoded training set. In the case of a binary symmetric channel, the channel is completely specified by a single parameter, ϵ, the channel bit error rate. The JSC decoder objective is to produce an approximation of the source $\underline{X}$, denoted $\underline{\hat{x}}^{(\mathrm{dec})} \equiv (\hat{x}_0^{(\mathrm{dec})}, \hat{x}_1^{(\mathrm{dec})}, \ldots, \hat{x}_{T-1}^{(\mathrm{dec})})$, given a realization of the noisy index sequence, $\underline{j} \equiv (j_0, j_1, \ldots, j_{T-1})$ [1]. The knowledge brought to bear by the decoder consists of the source and channel probability models.

The approach that we suggest builds on/*leverages* previous decoding techniques, such as [8]. Accordingly, we first review [8] and then show how our new approach both uses previous decoding results and, further, improves upon them.

Review of SAMMSE Decoder

The SAMMSE decoder[8] views the tandem of source and channel encoder as a hidden Markvo model. Thus, the decoder is based on estimation of the hidden state sequence (encoded source symbols) $\underline{I}$ given the the noisy symbol sequence $\underline{J}$. The decoder first finds the *a posteriori* probabilities $P[I_t = i_t | \underline{J} = \underline{j}] \quad \forall t$ by using the well known forward/backward algorithm[1]. These *a posteriori* probabilities are then used in a conditional mean estimator to reconstruct/estimate the quantization levels associated with each transmitted symbols. The derivation of the method is summarized as follows.

The overall distortion, caused by quantization error and decoder error, is given by $D(\underline{x}, \underline{\hat{x}}^{(\mathrm{dec})}) = \sum_{t=0}^{t=T-1} d(x_t, \hat{x}_t^{(\mathrm{dec})})$, where $d(x_t, \hat{x}_t^{(\mathrm{dec})}) = (x_t - \hat{x}_t^{(\mathrm{dec})})^2$. The decoder's objective is to minimize the expected distortion over $\underline{\hat{x}}^{(\mathrm{dec})}$ given the received noisy symbol sequence $\underline{j}$. This equation is given by

$$\begin{aligned} & E[D(\underline{x}, \underline{\hat{x}}^{(\mathrm{dec})}) | \underline{J} = \underline{j}] \qquad (2) \\ = \; & \frac{\sum_{\underline{i}} \Big(\sum_{t=0}^{t=T-1} E[d(x_t, \hat{x}_t^{(\mathrm{dec})}) | \underline{I} = \underline{i}] \Big) P[\underline{J} = \underline{j} | \underline{I} = \underline{i}] P[\underline{I} = \underline{i}]}{\sum_{\underline{i}} P[\underline{J} = \underline{j} | \underline{I} = \underline{i}] P[\underline{I} = \underline{i}]}. \end{aligned}$$

The authors in [8] have shown that directly working with this equation is not practical. Therefore, the approximation $E[d(x_t, \hat{x}_t^{(\mathrm{dec})}) | \underline{I} = \underline{i}] \approx E[d(x_t, \hat{x}_t^{(\mathrm{dec})}) | I_t = i_t]$ is needed. By using this approximation, (2) can be further reduced to minimizing

$$\sum_{t=0}^{t=T-1} \sum_{l=1}^{l=L} \| E[x_t | I_t = l] - \hat{x}_t^{(\mathrm{dec})} \|^2 P[I_t = l | \underline{J} = \underline{j}], \tag{3}$$

where

$$P[i_t = l | \underline{J} = \underline{j}] = \frac{\sum_{\underline{i}: i_t = l} p[\underline{J} = \underline{j} | \underline{I} = \underline{i}] P[\underline{I} = \underline{i}]}{\sum_{\underline{i}} P[\underline{J} = \underline{j} | \underline{I} = \underline{i}] P[\underline{I} = \underline{i}]}. \tag{4}$$

The probabilities $P[\underline{I} = \underline{i}]$ can be obtained assuming a low order Markov model as shown in equation (1). For a memoryless channel, the probabilities $P[\underline{J} = \underline{j} | \underline{I} = \underline{i}]$ can be simplified to $\prod_{t=0}^{t=T-1} P[J_t = j_t | I_t = i_t]$ with $P[J = j | I = i]$ assumed known. Minimizing (3) leads to choosing the reconstructed values $\hat{x}_t^{(\mathrm{dec})}$ according to the sequence-based approximation minimum mean squared error (SAMMSE) rule:

$$\hat{x}_t^{(\mathrm{dec})} = \sum_{l=1}^{l=L} E[x_t | I_t = l] P[I_t = l | \underline{J} = \underline{j}], \qquad \forall t. \tag{5}$$

[1] Without knowledge of $\underline{j}$, the decoder output $\hat{X}_t^{(\mathrm{dec})}$ is treated as a random variable. However, the decoding rule is a deterministic function of the received sequence $\underline{j}$. For purpose of retaining compact notation, we do not explicitly indicate the dependence on $\underline{j}$ in $\underline{\hat{x}}^{(\mathrm{dec})}$.

$P[I_t = l|\underline{J} = \underline{j}]$ can be obtained by using the forward/backward algorithm[1] with $E[x_t|I_t = l]$ again estimated from a large training set.

There are several ways to improve SAMMSE decoding. One strategy is to increase the order of the Markov model for $\{I_t\}$. However, the complexity of the forward and backward recursions needed to calculate the *a posteriori* probabilities grows exponentially with the Markov order K, i.e. the complexity is $O(N^{(K+1)}T)$. This limits the practical feasibility of this approach. A second potential strategy is to use "memory-enhanced" decoding, as suggested in [8]. This approach uses a higher resolution decoder lookup table and thus obtains a more accurate conditional mean estimate. For the case of second order (enhanced memory) decoding, the decoding rule is:

$$E[X_t|\underline{j}] = \sum_l \sum_m Q_{\mathrm{dec}}^{-1}(l,m) P[I_t = l, I_{t-1} = m|\underline{j}], \quad (6)$$

where $Q_{\mathrm{dec}}^{-1}(l,m) \equiv E[X_t|I_t = l, I_{t-1} = m]$. It should be emphasized that the decoder memory can be usefully chosen to be second or higher order *irrespective of* the order of the Markov model for $\{I_t\}$ [2]. Unfortunately, the number of summations increases linearly, and thus the number of summands *exponentially*, with the order of the decoder memory. Moreover, the complexity of calculating the *a posteriori* probabilities $P[I_t, I_{t-1}, \ldots, I_{t-K_d+1}|\underline{j}]$, *also* increases exponentially with the decoder memory.

While both increasing the Markov order and increasing the decoder memory improve performance, there is a heavy price in complexity. In the next section, alternatively, we will demonstrate a way of improving the SAMMSE decoder with only *modest* increases in complexity.

A New JSC Decoder

As just discussed, JSC decoding can be improved by increasing the Markov order or the decoder's memory, albeit with exponential growth in complexity. Alternatively, here we suggest a heuristic way to increase the decoder "order" with minimal growth in complexity. In particular, we suppose that if the decoder model is not fully capturing the memory in the source, then it may very well be the case that there is unexploited *partial correlation* between the source X_t and both causal $\{\hat{X}_{t-k}^{(\mathrm{dec})}, k > 0\}$ and *anticausal* $\{\hat{X}_{t+k}^{(\mathrm{dec})}, k > 0\}$ decoded values. In making this statement, we are recognizing that $\hat{X}_{t-k}^{(\mathrm{dec})}(\underline{J})$ is truly a *random* quantity, as it is a function of the random sequence $\underline{J}$. If there is untapped correlation, this suggests the possibility of improving the decoding result $\hat{X}_t^{(\mathrm{dec})}$ via linear filtering, i.e. forming

$$\hat{x}_t^{(\mathrm{newdec})} = \sum_{k=0}^{L_c-1} \alpha_k \hat{x}_{t-k}^{(\mathrm{dec})} + \sum_{k=1}^{L_{\mathrm{nc}}} \alpha_{-k} \hat{x}_{t+k}^{(\mathrm{dec})}. \quad (7)$$

(7) amounts to a *two-stage* estimation procedure, with standard SAMMSE decoding first applied(5), followed by linear filtering. Low order (e.g. $K = 1$) SAMMSE decoding followed by the filtering in (7) is *far* less complex than SAMMSE decoding based on a high order Markov model. The complexity comparison will be further discussed in the experimental result section. It remains to determine the coefficients $\{\{\alpha_k\}, \{\alpha_{-k}\}\}$.

In JSC decoding, e.g. [8, 12, 13], the approach often taken is to assume an optimality criterion (such as MMSE or MAP) and a statistical model and then to analytically derive a closed form decoder expression. Alternatively, we take our cue from *training-based* approaches to quantizer design, proposing a training-based approach to the design of a JSC decoder that captures residual redundancy in the source. The ultimate performance is the MSE $E[(X - \hat{X}^{(\mathrm{dec})})^2]$, with the expectation taken with respect to both the source and channel distributions. Now, as is often done in practice without proof, e.g. [5], let us suppose that an *ergodicity property* holds, i.e., in our case, that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (x_t - \hat{x}_t^{(\mathrm{dec})}(\underline{j}))^2 = E[(X - \hat{X}^{(\mathrm{dec})})^2], \quad (8)$$

where we have emphasized the rule's dependence on $\underline{j} = (j_0, j_1, \ldots, j_{T-1})$. The motivation behind this is that, if (8) holds, then one can choose the rule $\hat{x}_t^{(\mathrm{dec})}(\underline{j})$ to minimize a least squares cost based on a large training set *and a single realization of the channel*, $\underline{j}$, with the reasonable expectation that one is then (approximately) choosing the decoder to minimize the MSE $E[(X - \hat{X}^{(\mathrm{dec})})^2]$. A similar approach was taken in [3] for a different source coding context and for optimization at the *encoder*, not the decoder. The reasonableness of our ergodicity assumption will be substantiated by our experimental results. In particular, it will be seen that the decoding performance optimized over the training set closely agrees with the performance on multiple independent test sets.

Thus, the coefficients $\{\{\alpha_k\}, \{\alpha_{-k}\}\}$ can be chosen to minimize the *least squares error* (LSE) performance criterion,

$$\sum_{t=L_c-1}^{T-L_{nc}-1} (x_t - \hat{x}_t^{(\mathrm{dec})}(\underline{j}))^2. \quad (9)$$

The LS-optimal coefficients are given by the standard LS solution:

$$\underline{\alpha} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \underline{x}, \quad (10)$$

with

$$\mathcal{X}^T = \begin{bmatrix} \hat{x}_0^{(\mathrm{dec})} & \hat{x}_1^{(\mathrm{dec})} & \cdots & \hat{x}_{T-L_c-L_{nc}}^{(\mathrm{dec})} \\ \hat{x}_1^{(\mathrm{dec})} & \hat{x}_2^{(\mathrm{dec})} & \cdots & \hat{x}_{T-L_c-L_{nc}+1}^{(\mathrm{dec})} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{x}_{L_c+L_{nc}-1}^{(\mathrm{dec})} & \hat{x}_{L_c+L_{nc}}^{(\mathrm{dec})} & \cdots & \hat{x}_{T-1}^{(\mathrm{dec})} \end{bmatrix}, \quad (11)$$

$\underline{\alpha} = (\alpha_{L_c-1}, \alpha_{L_c-2}, \alpha_{L_c-3}, \cdots, \alpha_0, \alpha_{-1}, \cdots, \alpha_{-L_{nc}})^T$,and $\underline{x} = (x_{L_c-1}, x_{L_c}, \cdots, x_{T-L_{nc}-1})^T$ obtained from a training sequence. The resulting solution can be rewritten in the (explicit) form:

$$\begin{aligned} \hat{x}_t^{(\mathrm{newdec})} &= \sum_{k=0}^{L_c-1} \alpha_k (\sum_{l=0}^{N} P[I_{t-k} = l|\underline{j}] y(l)) \\ &+ \sum_{k=1}^{L_{\mathrm{nc}}} \alpha_{-k} (\sum_{l=0}^{N} P[I_{t+k} = l|\underline{j}] y(l)), \quad (12) \end{aligned}$$

where $y(l) \equiv E[X|I = l]$.

This form is suggestive of a way to further *improve* the solution. In particular, the quantities $y(l) \equiv E[X|I = l]$ are usually estimated based on an encoded training set. Alternatively, we can replace the products $\alpha_k y(l)$ by *parameters* $\mu_{k,l}$,

[2] It is also not necessary for the conditioning context to be causal. It could also be noncausal, e.g. $E[X_t|I_t = m, I_{t-1} = l, I_{t+1} = n]$.

with the decoding rule now

$$\begin{aligned}\hat{x}_t^{(\mathrm{newdec})} &= \sum_{k=0}^{L_c-1}\sum_{l=0}^{N-1} \mu_{k,l} P[I_{t-k}=l|\underline{j}] \\ &+ \sum_{k=1}^{L_{\mathrm{nc}}}\sum_{l=0}^{N-1} \mu_{-k,l} P[I_{t+k}=l|\underline{j}].\end{aligned} \tag{13}$$

The $\{\mu_{k,l}\}$ can again be chosen to minimize the least squares cost (9). The solution is again given by the form (10), but this time with the data matrix defined as

$$\mathcal{X}^T = \left[\begin{array}{cccc} \underline{P[I_0|\underline{j}]} & \underline{P[I_1|\underline{j}]} & \cdots & \underline{P[I_{T-L_c-L_{nc}}|\underline{j}]} \\ \underline{P[I_1|\underline{j}]} & \underline{P[I_2|\underline{j}]} & \cdots & \underline{P[I_{T-L_c-L_{nc}+1}|\underline{j}]} \\ \vdots & \vdots & \vdots & \vdots \\ \underline{P[I_{L_c+L_{nc}-1}|\underline{j}]} & \underline{P[I_{L_c+L_{nc}}|\underline{j}]} & \cdots & \underline{P[I_{T-1}|\underline{j}]} \end{array}\right], \tag{14}$$

where

$$\underline{P[I_t|\underline{j}]} = (P[I_t=0|\underline{j}], P[I_t=1|\underline{j}], \ldots, P[I_t=N-1|\underline{j}])^T,$$

and

$$\begin{aligned}\underline{\alpha} = (&\mu_{L_c-1,0}, \mu_{L_c-1,1}, \ldots, \mu_{L_c-1,N-1}, \ldots, \mu_{0,0}, \ldots, \mu_{0,N-1}, \\ &\mu_{-1,0}, \ldots, \mu_{-1,N-1}, \ldots, \mu_{-L_{nc},0}, \ldots, \mu_{-L_{nc},N-1})^T.\end{aligned}$$

The performance of both (12) and (13) are evaluated in our experimental results. Next, we briefly discuss the implications of our LS approach for *predictive* JSC decoding.

JSC Decoding for Predictively Encoded Sources

In [4, 6, 10, 13], a common JSC decoding strategy for predictively encoded sources (e.g. DPCM) is to first form a JSC decoding estimate of the prediction residual and then feed this estimate to a standard (noise-free) DPCM decoder. For example, in [10], for the case of first order DPCM, the proposed decoding rule was:

$$\hat{x}_t^{(\mathrm{dec})} = a\hat{x}_{t-1}^{(\mathrm{dec})} + E[Z_t|\underline{j}], \tag{16}$$

where $E[Z_t|\underline{j}] = \sum_{l=0}^{N-1} c(l)P[I_t=l|\underline{j}]$ and $c(l) \equiv E[Z_t|I_t=l]$. Here, Z_t is the prediction residual and a is the prediction coefficient. Note that $E[Z_t|\underline{j}]$ is just the SAMMSE decoding estimate of the prediction residual.

The main principle behind our LS decoding strategy is to treat the standard JSC decoding outputs $\hat{X}_t^{(\mathrm{dec})}(\underline{J})$ as *data*, i.e., as random observations, correlated with the true signal X_t. Accordingly, in the last section we designed linear filters to provide an LS estimate of X_t given $\{\hat{X}_t^{(\mathrm{dec})}(\underline{J})\}$. This approach improves performance at the cost of *some* increase in implementation complexity, associated with the linear filtering. However, the complexity increase is modest compared with increasing the Markov order or using memory-enhanced decoding[8]. In the predictive coding case, use of our LS strategy can yield improved decoders with *no* complexity increase, i.e. a fundamental improvement can be achieved using our LS decoding strategy. Alternatively, further gains can again be achieved with modest increase in complexity. Our LS design strategy for JSC decoding of predictively-encoded sources was proposed in [7]. Here we summarize [7]. The key observation is that the predictive JSC decoding rule (16) is *already* in the form of a linear estimator, based on the "observations" $\hat{x}_{t-1}^{(\mathrm{dec})}$ and $E[Z_t|\underline{j}]$. This raises the question of whether the coefficients $(a, 1)$ are *optimal* (or nearly so) in the least squares sense, as the weights for these observations ?

Consider the more general estimator

$$\hat{x}_t^{(\mathrm{newdec})} = \alpha\hat{x}_{t-1}^{(\mathrm{dec})} + \beta E[Z_t|\underline{j}], \tag{17}$$

with $\hat{x}_{t-1}^{(\mathrm{dec})}$ obtained from the standard decoding rule (16). I.e., suppose (16) is first applied, yielding $\{\hat{x}_t^{(\mathrm{dec})}\}$, with the optimal coefficients then sought to weight the values $\hat{x}_{t-1}^{(\mathrm{dec})}$ and $E[Z_t|\underline{j}]$ in forming a *new* decoding estimate $\hat{x}_t^{(\mathrm{newdec})}$. For a training set $\underline{x} = (x_0, x_1, \ldots, x_{T-1})^T$ and a realization of the channel $\underline{j}$, the pair (α, β) optimal in the least squares sense again satisfies

$$(\alpha\ \beta)^T = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\underline{x}, \tag{18}$$

with

$$\mathcal{X}^T = \left[\begin{array}{cccc} \hat{x}_{-1}^{(\mathrm{dec})} & \hat{x}_0^{(\mathrm{dec})} & \cdots & \hat{x}_{T-2}^{(\mathrm{dec})} \\ E[Z_0|\underline{j}] & E[Z_1|\underline{j}] & \cdots & E[Z_{T-1}|\underline{j}] \end{array}\right], \tag{19}$$

and with $\hat{x}_{-1}^{(\mathrm{dec})}$ an initial value. Our approach to JSC decoding for predictively encoded sources[7], similar to the approach for nonpredictive coding, is to use the LS-optimal coefficients in the decoding rule (17).

Consider an example of a first order Gauss-Markov process $\{X_t\}$ with correlation coefficient 0.95, $N = 8$ quantization levels, a first order Markov model for the sequence $\{I_t\}$, and a binary symmetric channel with bit error rate $\epsilon = 0.05$. Residual redundancy is introduced by mismatching the prediction coefficient, relative to the source correlation. Suppose the prediction coefficient is chosen as $a = 0.45$. Experimentally, we have found that this choice leads to good decoding performance both for our system and for the standard JSC decoder (which uses $(a, 1)$). For this choice, we find that the LS-optimal pair (based on a training set of 10^6 samples), is $(\alpha, \beta) = (0.55, 0.79)$, quite different from the values $(0.45, 1)$ that give the standard rule. Moreover, averaged over 3 test sets of size fifty thousand samples, the reduction in distortion of this LS-optimal decoder over the standard JSC decoder is 10log 10(MSE of standard JSC/MSE of new JSC) = 0.41 dB. Clearly, the standard JSC rule is *not* in general the optimal way to combine the "observations" $E[Z_t|\underline{j}]$ and $\hat{x}_{t-1}^{(\mathrm{dec})}$.

Even better performance is achieved by increasing the order of the filter, i.e. forming:

$$\hat{x}_t^{(\mathrm{newdec})} = \sum_{k=1}^{L_c} \alpha_k \hat{x}_{t-k}^{(\mathrm{dec})} + \beta E[Z_t|\underline{j}], \ L_c \geq 1. \tag{20}$$

Moreover, while (20) only includes causal terms, our method can also be applied to design a decoder that uses *anticausal* terms $\{\hat{x}_{t+k}^{(\mathrm{dec})}, k \geq 0\}$, with the potential for additional gains in performance. In the next section, we evaluate the LS design approach we have developed here (and the various nonpredictive and predictive decoding structures) in comparison with conventional JSC decoding approaches.

III. Experimental Results

We investigated our LS optimization approach for improving the performance of first order SAMMSE decoding. As a source, we chose the first AC (AC1) DCT coefficient from 8x8 blocks of a gray scale image. This one dimensional AC1 source was created by 1) DCT transforming the image using 8x8 blocks, 2)collecting the same transform coefficients from each of the 8x8 block to form 64 different coefficient "images", 3)for each image, a 1D source is created by raster scanning; We chose to encode the source with second highest variance (AC1). As a source encoder, we used a scalar quantizer with eight quantization levels (3 bit fixed length). This quantizer was designed via the Lloyd algorithm using the AC1 source extracted from 23 gray scale images. We chose the bit error rate of our binary symmetric channel to be 0.05. Both the first and second order SAMMSE decoders were designed based on these images. The Markov model probabilities for these decoders were learned based on frequency counts from the AC1 coefficients for all 23 images. The first order SAMMSE decoder gave an average SQNR($10log_{10}\frac{\sigma_x^2}{MSE}$) of 2.91 dB over these 23 image source. The second order SAMMSE decoder based on the same source, gave an average SQNR of 4.52 dB.

For our method, we first designed an LS-optimal filter for the first order SAMMSE decoding result based on the decoder form (12). Our decoding performance was tested on all 23 images with LS filter coefficients optimized for each image [3]. The result is shown in Figure 2. In this Figure, the 'number of filter' taps include both causal and anticausal coefficients. The number of filter taps increases by adding one tap from causal and one tap from anticausal context alternately, initialized by one causal filter tap. Let us consider the case when $L_c = 3$ and $L_{nc} = 2$. With these filter taps, only a small amount of side information is needed to specify the filter coefficients to the decoder for each image. The resulting averaged LS-optimal performance was 4.12 dB. This is approximately 1.2 dB better than using only a first order SAMMSE decoder, and is only approximately 0.4 dB worse than a second order SAMMSE decoder. In this case, the LS decoder complexity is roughly 8 % higher than first order SAMMSE and roughly 7.4 times less complex than second order SAMMSE.

We also designed an LS-optimal filter that filters the *a posterior* probabilities of a first order SAMMSE decoding result based on the decoder form (13). The result is shown in Figure 3. In this figure, the number of filter taps also include both causal and anticausal coefficients. Each "tap" in this case corresponds to eight coefficients. These 8 coefficients are associated with eight possible decoded symbol values at each symbol time. The number of filter tap sets $\underline{\mu_k}$ increases in the same fashion as in Figure 2. Let us note the case when there is only one set of filter taps. The resulting averaged LS-optimal performance was 4.41 dB. This is approximately 1.5 dB better than using only a first order SAMMSE decoder, and is only approximately 0.1 dB worse than a second order SAMMSE decoder. In this case, the LS decoder complexity is roughly 12.5 % higher than first order SAMMSE and roughly 7.1 times less complex than second order SAMMSE. Compared to the same computational complexity with 8 filter taps in Figure 2, this approach is approximately 0.3 dB better.

In Table 1, a comparison between the standard predictive decoder(16) and the LS optimized predictive decoder(20) using 10 causal coefficients is shown. This predictive result was reported in [7]. In this experiment, we used a first order Gauss-Markov source $\{X_t\}$ with correlation coefficient $\rho = 0.95$ and with input white noise variance $\sigma_w^2 = 1.0$. A training set of 1 million samples was used for this decoder design. Three independent test sets were generated, each of size 50,000 samples. We assumed first order DPCM at 3 bits/sample. A uniform quantizer was chosen, based on the dynamic range of the prediction residual. For SAMMSE estimation of the prediction residual $E[Z_t|\underline{j}]$, we assumed a first order Markov model for $\{I_t\}$. This table shows the performance gain of using LS optimization approach in the predictive case with different choices of prediction coefficient.

[3]We found that only small performance gain was achieved if a single "universal" filter $\{\alpha\}$ is used for all the images

IV. Conclusions

Although higher order modeling can improve decoding performance, the direct implementation of high order (>3) SAMMSE decoding is typically too complex for modern computers. In this work, we proposed a method to bridge the performance-complexity gap between direct implementation of low and high order SAMMSE decoding. We showed that our LS filtering approach was able to increase decoding performance with a small increase in computational complexity. Our LS filter is chosen to provide an estimate of the original source via a large training set. This approach provides a way to approximate the effective order increase in the standard approach without the exponential increase in computational complexity. We have verified our approach by first designing a LS filter for the decoding form (12). We then further improved this approach by designing the LS filter for the form (13). The LS filtering approach was also applied to predictively encoded sources with a significant decoding performance gain.

Figure 1: The basic communication system

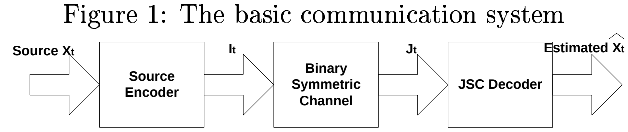
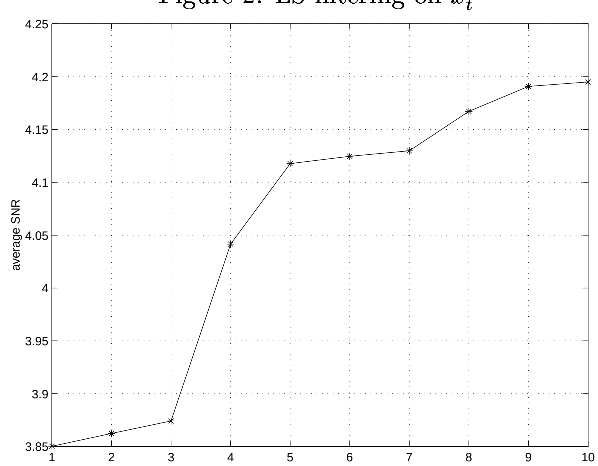


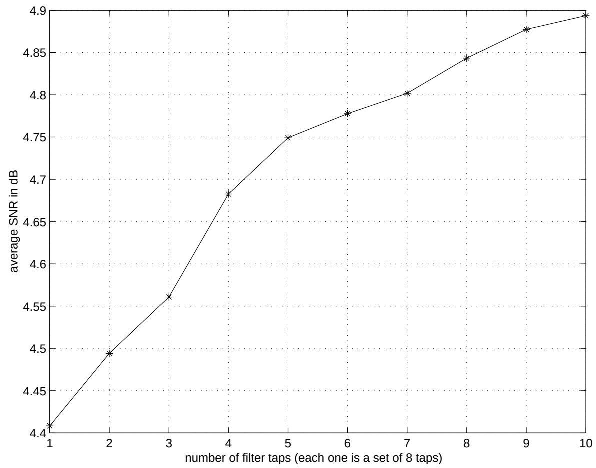
Figure 2: LS filtering on $\hat{x}_t^{(\mathrm{dec})}$



References

[1] L. Baum, T. Petrie, G. Soules, and N. Weiss. A Maximization Technique Occurring in the Statistical Analysis

Figure 3: LS filtering on *a posteriori* prob. $p[I_t = l|\underline{j}]$



Prediction coefficient	Std. Dec. Avg. SQNR	LS Dec. Avg. SQNR	Gain vs Std. Dec.
0.35	11.761	12.533	0.772
0.45	12.008	12.539	0.531
0.55	11.987	12.354	0.367

Table 1: SQNR performance of standard and LS decoding for DPCM encoding of a Gauss-Markov source, as a function of the prediction coefficient. In this case, the LS decoder uses 10 causal samples.

of Probabilistic Functions of Markov Chains. *Annals of Math. and Stat.*, 41:164–171, 1970.

[2] P. Bunyaratavej and D. Miller. An Iterative Hillclimbing Algorithm for Discrete Optimization on Images: Application to Joint Encoding of Image Transform Coefficients. *IEEE Signal Processing Letters*, 9(2):46–50, Feb. 2002.

[3] K.-H. Chei and K.-P. Ho. Design of Optimal Soft Decoding for Combined Trellis Coded Quantization/Modulation in Rayleigh Fading Channel. *IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc.*, pages 2633–2636, 2000.

[4] S. Emami and S. Miller. DPCM Picture Transmission over Noisy Channel with the Aid of a Markov Model. *IEEE Transactions on Image Processing*, 4(11):1473–1479, Nov. 1995.

[5] J. D. Jr., J. Proakis, and J. Hansen. *Discrete-time Processing of Speech Signals.* Macmillan Publishing, New York, 1993.

[6] R. Link and S. Kallel. Markov Model Aided Decoding for Image Transmission using Soft-Decision-Feedback. *IEEE Transactions on Image Processing*, 9(2):190–196, Feb. 2000.

[7] D. Miller, E. S. Carotti, Y. Wang, and J. C. de Martin. Joint Source-Channel Decoding of Predictively and Non-Predictively Encoded Sources: a Two-Stage Estimation Approach. *IEEE Transactions on Communications.* Submitted to.

[8] D. Miller and M. Park. A Sequence-based Approximate MMSE Decoder for Source Coding over Noisy Channels Using Discrete Hidden Markov Models. *IEEE Transactions on Communications*, 46(2):222–231, Feb. 1998.

[9] A. Murad and T. Fuja. Joint Source-Channel Decoding of Variable-Length Encoded Sources. *Information Theory Workshop*, (22–26):94–95, June 1998.

[10] M. Park and D. Miller. Improved Image Decoding over Noisy Channels using Minimum Mean-Squared Estimation and a Markov Mesh. *IEEE Transactions on Image Processing*, 8(6):863–867, June 1999.

[11] M. Park and D. Miller. Joint Source-Channel Decoding for Variable-Length Encoded Data by Exact and Approximate MAP Sequence Estimation. *IEEE Transactions on Communications*, 48:1–6, Jan. 2000.

[12] N. Phamdo and N. Farvardin. Optimal Detection of Discrete Markov Sources over Discrete Memoryless Channels-Application to Combined Source-Channel Coding. *IEEE Transactions on Inform. Theory*, 40:186–193, 1994.

[13] K. Sayood and J. Borkenhagen. Use of Residual Redundancy in the Design of Joint Source/Channel Coders. *IEEE Transactions on Communications*, 39(6):838–846, June 1991.